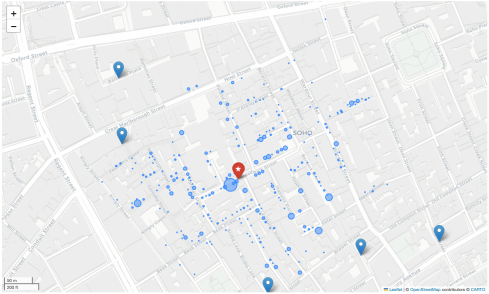
Heat Maps and Kernel Density Estimation

Jonathan Taylor
Stanford University
DATASCI 112

June 1, 2026

1. From Dot Maps to Heat Maps
2. Kernel Density Estimation (KDE)
3. Connection to Classification

Limitations of Dot Maps



Recall John Snow’s dot map of cholera cases.

Limitations of Dot Maps

Dot maps are great, but they have limitations:

- **Overplotting:** If many points are in the same location, they overlap and look like a single dot.
- **Discrete vs. Continuous:** Points are discrete, but we often want to understand the underlying continuous aspect of a phenomenon.

Heat Maps

A **heat map** represents the “density” of points using color.

Intensity of color corresponds to the number of points in an area. Besides visualization, “density” has many applications in data science.

1. From Dot Maps to Heat Maps
2. Kernel Density Estimation (KDE)
3. Connection to Classification

What is KDE?

Kernel Density Estimation (KDE) is the mathematical technique used to create heat maps.

Intuition: Place a small "bump" (a kernel) on top of every data point. Sum up all the bumps to get the final "density".

Density: A non-negative function that represents "chances": area under the density totals 1:

$$\int f(x)\,dx = 1$$

Makes sense even in dimension $d > 1$:

$$\hat{f}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

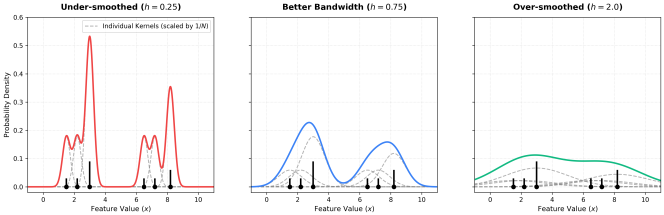
- n: number of points.
- h: **bandwidth** (controls the width of the bumps).
- K: **kernel** function (e.g., Gaussian).

Bandwidth Choice

The bandwidth h is a critical parameter.

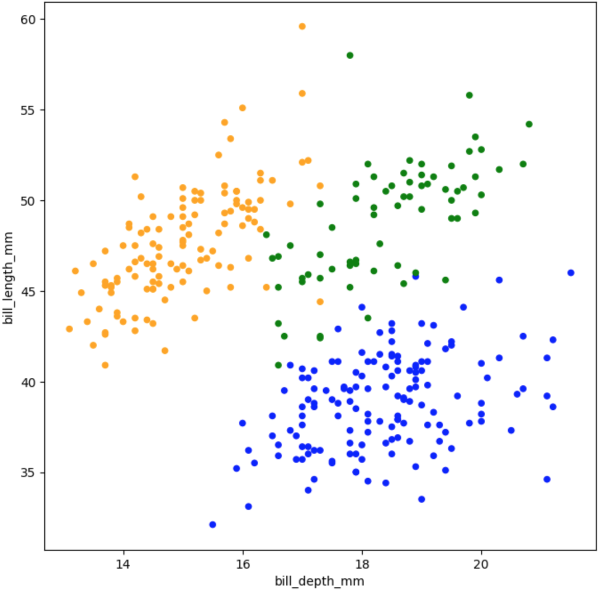
- **Small h:** The density is very wiggly and peaks at every data point (overfitting).
- **Large h:** The density is very smooth and loses detail (underfitting).
- **Choice of h:** Software packages have pretty reliable choices of bandwidth...

Choice of bandwidth

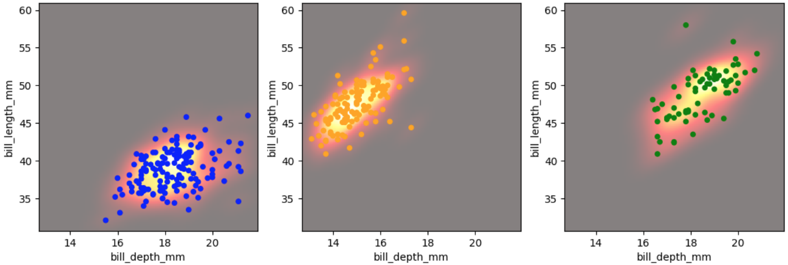


In the classification setting, we can define a “density” for each class.

Penguin data



Densities for each penguin class



A heuristic: if density for `Gentoo` is high at `(depth,length)`, it might be `Gentoo`

We should compare to density of `Chinstrap` and `Adelie` at `(depth,length)` as well.

KDE for Classification

We can perform classification by estimating the density $\hat{f}_k(x)$ for each class k separately.

For each x and class k, assign a score

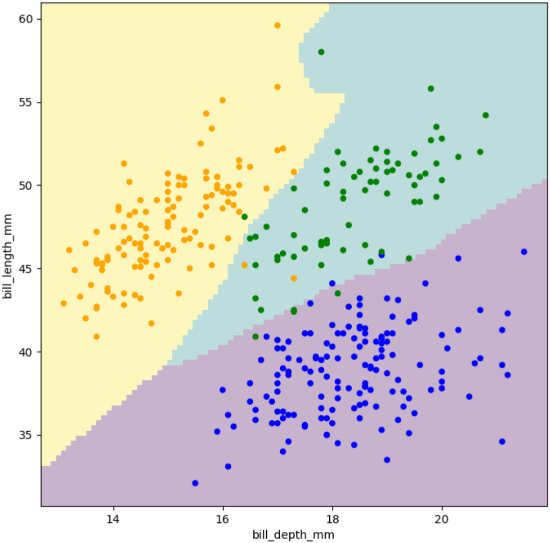
$$S_k(x) = \hat{\pi}_k \hat{f}_k(x)$$

- π_k: **Prevalence** (prior probability) of class k – controls for class imbalance.
- $\hat{f}_k(x)$: KDE estimate for class k (or other density estimate...)

Classification rule: Assign x to the class k which maximizes S_k:

$$\text{Class}(x) = \text{argmax}_{1 \leq k \leq K} S_k(x).$$

Densities for each penguin class



Relation to LDA

Recall **Linear Discriminant Analysis (LDA)**.

LDA is exactly this model, but with a specific assumption:

- Each class density $f_k(x)$ is a **single Multivariate Gaussian**: one *big* bump whose shape is determined by an ellipse...
- All classes share the **same covariance matrix** – this covariance matrix controls the ellipse.

KDE-based classification is more flexible because it can model non-Gaussian, multi-modal densities.

Compare to LDA boundaries